BMJ Open

Confounding adjustment methods in longitudinal observational data with a time-varying treatment: a mapping review

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-058977
Article Type:	Original research
Date Submitted by the Author:	10-Nov-2021
Complete List of Authors:	Wijn, Stan; Radboudumc, Radboud university medical center, Radboud Institute for Health Sciences, Department of Operating Rooms Rovers, Maroeska; Radboudumc, Radboud university medical center, Radboud Institute for Health Sciences, Department of Operating Rooms and Health Evidence Hannink, Gerjon; Radboudumc, Radboud university medical center, Radboud Institute for Health Sciences, Department of Operating Rooms
Keywords:	EPIDEMIOLOGY, STATISTICS & RESEARCH METHODS, Orthopaedic & trauma surgery < SURGERY

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Confounding adjustment methods in longitudinal observational data with a time-varying treatment: a mapping review

Stan R.W. Wijn[1], Maroeska M. Rovers[1,2], Gerjon Hannink[1]

[1] Radboud University Medical Centre, Radboud Institute for Health Sciences, Department of Operating Rooms, Nijmegen, the Netherlands

[2] Radboud University Medical Centre, Radboud Institute for Health Sciences, Department of Health Evidence, Nijmegen, the Netherlands

S.R.W. Wijn, Stan.Wijn@radboudumc.nl

M.M. Rovers, Maroeska.Rovers@radboudumc.nl

G. Hannink, Gerjon.Hannink@radboudumc.nl

Corresponding author:

Stan R.W. Wijn
Radboud university medical centre
715 Department of Operating Rooms
P.O. Box 9101
6500 HB Nijmegen
The Netherlands

Declarations of interest: none

Word count: 2114

Keywords: Propensity score matching, longitudinal observational data, time-varying treatment, confounding, g-methods

Abstract

Objectives

To adjust for confounding in observational data, researchers use propensity score matching (PSM), but more advanced methods might be required when dealing with longitudinal data and time-varying treatments as PSM might not include possible changes that occurred over time. This study aims to explore which confounding adjustment methods have been used in longitudinal observational data to estimate a treatment effect and identify potential inappropriate use of PSM.

Design

Mapping review

Study Design and Setting

We searched PubMed for papers in which a treatment was evaluated using longitudinal observational data from inception up to January 2021. Methodological-, non-medical-, and cost-effectiveness papers were excluded as well as studies that did not study a treatment effect. They were categorized based on time of treatment: at baseline (interventions performed at a start of follow-up) or time-varying (interventions received asynchronous during follow-up). Studies were sorted based on publication year, time of treatment and confounding adjustment method. Cumulative time series plots were used to investigate the use of different methods over time. No risk of bias assessment was performed as it was not applicable.

Results

In total, 760 studies were included that met the eligibility criteria. PSM (165/201, 82%) and inverse probability weighting (150/498, 30%) were most common for studies with a treatment at baseline (n=201) and time-varying treatment (n=498), respectively. Of the 498 studies with a time-varying treatment, 123 (25%) used PSM with baseline covariates, which might be inappropriate. In the last

five years, the proportion of studies with a time-varying treatment that used PSM over IPW increased.

Conclusions

PSM is the most frequently used method to correct for confounding in longitudinal observational data. In studies with a time-varying treatment PSM was potentially inappropriate in 25% of the studies. Confounding adjustment methods designed to deal with a time-varying treatment and time-varying confounding are available, but are not regularly used.

Article Summary

Strengths and limitations of this study

- We systematically mapped the literature for the most commonly used methods to correct for confounding in longitudinal observational data.
- Although time-dependent methods like time-dependant propensity score matching, parametric g-formula and inverse probability weighting are described in detail in the literature, adjusting at baseline in observational data is still common and was potentially inappropriate in a proportion of the papers we included in our mapping review.
- A limitation of a mapping review is the broad descriptive level at which studies are analysed. However, it does provide a general overview of the published literature.
- For some studies we were not able to identify if patients were treated at baseline or during follow-up. Fortunately, this only occurred in 8% of the included papers.

Introduction

The increasing availability of real-world data derived from electronic health records, registries, wearables, and surveys can be a valuable source of data to evaluate the effectiveness of a treatment.[1] Deriving inference directly from real-world data can be challenging as it is prone to confounding. To adjust for confounding, researchers use methods such as propensity score matching (PSM) to create two comparable groups in which both the treated- and untreated patients have similar observable characteristics (like age, pain scores, weight etc.) similar to a randomised trial.[2] Although these methods can be sufficient when a patient is treated at the start of a study (baseline), more advanced methods might be required when dealing with longitudinal data and time-varying or repeated treatments. Adjustment at baseline in the presence of longitudinal data and time-varying treatment might not include possible changes that occurred over time. These can include changes in treatment regimens or disease progression, but can also comprise weight changes, pain scores or changes in behaviour (e.g., stopped smoking). These changes can alter the balance between treated- and untreated patients and can result in different estimates of the treatment effect (see box 1).[3,4] Methods like time-dependent propensity score matching and the g-methods (inverse probability weighting (IPW), parametric g-formula or g-estimation) can incorporate time-varying covariates and time-varying treatments and can take feedback between the treatment and outcome over time into account.[2,5–8] It is however unclear if these methods are regularly used in practice when dealing with longitudinal observational data with a time-varying treatment. Therefore, this mapping review aimed to identify and describe which methods have been used to adjust for confounding bias in longitudinal observational data and identify potential inappropriate use of baseline adjustment methods (like PSM).

Box 1: Empirical example using data from the Osteoarthritis Initiative

To investigate the influence of the different confounding adjustment methods on the outcome, two previously published empirical examples with a time-varying treatment were selected: 1) the effect of meniscectomy (surgical removal of the meniscus) and 2) the effect of intra-articular corticosteroid injections on the risk to receive knee replacement surgery.[19,20] Data from the Osteoarthritis Initiative (OAI) was used for both examples. The OAI is a multicentre, longitudinal cohort study that included patients with (or at risk for) symptomatic femoral-tibial knee osteoarthritis (OA) with a follow-up up to 108 months, available for public access at https://data-archive.nimh.nih.gov/oai/. A large set of variables was extracted from the OAI, measured at baseline and annual follow-up visits. These include general patient characteristics, clinical variables, quality of life measurements, functional scores and time-varying treatments.

In total, we compared nine commonly used adjustment for both empirical examples: four methods that corrected using baseline covariates, four time-dependent methods, and no matching. We found in the first example (meniscectomy) that adjustment using baseline covariates resulted in larger estimates of the treatment effect compared to time-dependent methods, while results were consistent in the second example (intra-articular corticosteroid injection).(figure 1) These results show that the selected adjustment method can influence the detected treatment effect when dealing with potential time-varying confounding. See Supplement S2 for more details.

<insert figure 1>

Figure 1: Forest plot displaying the results of the two empirical examples (left: meniscectomy, right: intra-articular corticosteroid (IAC)). Four methods were compared using baseline covariates, four methods using time-dependent covariates and time-varying treatment and one without correction. PSM, propensity score matching; IPW, inverse probability weighting; CCA, conventional covariate adjustment; IAC, intra-articular corticosteroids; tdPSM, time-dependent propensity score matching.

Methods

A mapping literature review was performed to determine which confounding adjustment methods were used in longitudinal observational data to estimate a treatment effect. Mapping reviews are designed to map out and categorize existing literature and explore trends and identify gaps by study design and other key features.[9] This study was conducted and reported according to The PRISMA extension for Scoping Reviews (PRISMA-ScR).[10]

Patient and public involvement

Patients and/or public were not involved.

Search strategy

We searched in PubMed from inception up to January 2021 for papers in which a treatment was evaluated using longitudinal observational data. Search terms used were time varying, longitudinal observational data, and commonly used adjustment methods and terms (e.g., matching, g-methods). The search strategy can be found in Supplement SI. Methodological-, non-medical- and cost-effectiveness papers were excluded as well as non-English studies or studies that did not study a treatment effect. Studies that used no adjustment method or used the adjustment method solely as sensitivity analysis were also excluded. Study selection was performed by one reviewer and issues were discussed and resolved by all authors.

All papers were screened based on title and abstract and papers that met the inclusion criteria were screened full-text. The title, author(s), journal, research theme, publication date, confounding adjustment method, and time of treatment (at baseline or time-varying) were extracted from all papers that met the inclusion criteria. A treatment at baseline was defined as an intervention performed at the start of follow-up for all included patients (e.g., all treated patients received surgery at the start of follow-up). Time-varying treatment was defined as a treatment that was received asynchronous during follow-up (e.g., patients received surgery at different moments during follow-up) or when dealing with a repeated treatment of which the timing was not identical for all

treated patients (e.g., personalized medication intake over time). If the time of treatment was not defined, studies were categorized as unclear. No risk of bias assessment was performed because the scope of this paper targets the statistical methods that have been used in these papers, and therefore a risk of bias assessment was not applicable.

Analysis

Study selection was performed in Rayyan.[11] Study characteristics (author, publication year, journal), time of treatment (at baseline, time-varying or unclear) and confounding adjustment method were extracted and analysed in R (version 4.1.0, The R Foundation for Statistical Computing, Vienna, Austria). Studies were sorted based on publication year, time of treatment and confounding adjustment method and described using descriptive statistics. Cumulative time series plots were used to investigate the use of different methods over time for treatments at baseline and time-varying treatments using the Plotly package.[12]

Results

Our search identified 2134 articles of which eventually 760 met the eligibility criteria after title and abstract review, and subsequent full-text review (see also figure 2). The main reasons for exclusion were the lack of intervention/treatment (n = 405), a scope outside of medicine (n = 376), a methodological paper (n = 348), or the study did not utilize longitudinal observational data or did not correct for confounding (n = 123). Of all included papers, 201 (26%) had a treatment at baseline, 498 (66%) had a time-varying treatment and 61 (8%) papers had no clearly defined time of treatment. Of the papers with a treatment at baseline, the majority used PSM with baseline covariates (n = 165, 82%) as a method to correct for confounding. Studies that had a time-varying treatment most often used IPW (150 papers, 30%), PSM with baseline covariates was used in 123 papers (25%), PSM with baseline covariates combined with time-dependent Cox regression in 69 papers (14%), covariate adjustment using the propensity score in 49 papers (10%), time-dependent PSM in 40 papers (8%), parametric G-formula in 22 papers (4%), propensity score stratification in 18 papers (2%) and G-estimation in 13 papers (3%). In the last five years, the proportion of studies with a time-varying treatment that used PSM with baseline covariates over IPW increased (199 vs 158 in 2020, for PSM with baseline covariates and IPW, respectively). (Figure 3) For papers of which the time of treatment was unclear, PSM at baseline was most frequently used in 28 papers (46%).

<insert Figure 2>

Figure 2: PRISMA Flow Diagram of the flow of papers in the mapping review. In total, 760 studies were included and categorized according to the time of treatment. PSM, propensity score matching; IPW, inverse probability weighting; CA, covariate adjustment; PS, propensity score; TdPSM, time-dependent propensity score matching.

<insert Figure 3>

Figure 3: Cumulative incidence of the different confounding adjustment methods that are used in practice. Some studies used multiple methods. PSM, propensity score matching; IPW, inverse probability weighting; CA, covariate adjustment; PS, propensity score; TdPSM, time-dependent propensity score matching; PSS, propensity score stratification; RF, random forest matching.

Discussion

Although advanced methods are available to correct for confounding in longitudinal observational data, we showed that these methods are not always utilized in studies that have a time-varying treatment. Instead, 25% of the studies that had a time-varying treatment used PSM with baseline covariates to correct for confounding which can potentially result in a biased treatment effect.[4]

Our findings confirm the results by Clair et al. whom provided a summary of new methods that have been used in literature to deal with time-varying confounding. They concluded that IPW was most commonly used, more robust methods (like g-estimation) were underused.[13] Our results are also in agreement with the findings by Austin et al. whom reported a rapidly increasing use of IPW in the literature in the last decade.[14] Nonetheless, we detected a similarly rapid growth in the use of PSM in studies with a time-varying treatment, which can potentially result in biased results as PSM does not correct for time-varying confounding. Although time-dependent methods like tdPSM, parametric g-formula and IPW are extensively described in the literature [5,8,15], adjusting at baseline in observational data is still common in literature and was used in 25% of the papers with a time-varying treatment we included in our mapping review.[16] The proportion of studies with a time-varying treatment that used PSM over IPW even increased in the last five years.

Some potential limitations should also be discussed. First, the main limitation of a mapping review is the broad descriptive level at which studies are analysed and described. However, it does provide a general overview of the published literature and suggests that methods to deal with confounding in studies with a time-varying treatment are underused. Furthermore, no quality assessment of the included studies was performed. Second, although it is common to search multiple databases in a systematic review, our mapping review was limited to PubMed. We found over 2000 papers in Pubmed which was ample for the aim of this study and for a mapping review. It is unlikely that additional searches could alter our conclusions. Third, for some studies we were not able to identify

if patients were treated at baseline or during follow-up. Fortunately, this only occurred in 8% of the papers we included.

Implications

From previously published studies we can conclude that time-dependent methods can be important to avoid biased estimates of the treatment effect when adjusting for confounding in longitudinal observational data with potential time-varying confounding.[4,17] Therefore, we suggest using one of the g-methods (IPW, parametric g-formula, g-estimation) with time-varying covariates and time-varying treatment if the data is available.[17] Yet, these methods are not the panacea for unconfounded analyses in longitudinal observational data. They still rely on relevant confounder selection (based on prior knowledge, possibly supported by a directed acyclic graph), require careful examination of weights and adequate covariate balance.[14] Although there are clear benefits and limitations to each g-method, it is often unclear what the most appropriate method is to correct for confounding.[15] From the g-methods, IPW has three main advantages over the other methods: 1) it is a commonly used method, 2) it is relatively simple to understand and explain, and 3) it is easy to perform in standard statistical software (like R or STATA). Parametric g-formula is ideal for joint interventions or dynamic interventions but requires more computational power and additional programming.[17] G-estimation is particularly useful for studying the interaction between treatment and time-varying confounders (treatment-confounder feedback), but it can be challenging to implement g-estimation in longitudinal data. G-estimation can also be complex as there are not many practical guidelines or statistical packages that support this method for longitudinal data with a time-varying treatment. The developers of *gesttools* R-package (General Purpose G estimation in R) are currently drafting a comprehensive introduction including an explanation of the structural nested mean model types, the g-estimation algorithm, instructions to set up the users' dataset, and a tutorial to perform g-estimation.[18]

When dealing with real-world data, g-methods are recommended to evaluate the effectiveness of a treatment to preclude confounding. However, a proper assessment of the required confounding adjustment methods prior to data analysis is appropriate.

Conclusion

PSM using baseline covariates is the most used method to correct for confounding in longitudinal observational data, even in the presence of a time-varying treatment. Of the 498 identified studies with a time-varying treatment, 123 (25%) used PSM with baseline covariates, which might be inappropriate. Confounding adjustment methods designed to deal with a time-varying treatment and time-varying confounding are available, but are not regularly used and can potentially result in biased estimates of the treatment effect.

Declarations

Competing interests

The authors declare that they have no competing interest.

Data availability statement

Search strategies and data extraction documents are available on request to the corresponding author.

Ethics approval statement

This study does not involve human participants

Funding

This work was supported by the Junior Research project (2018) grant provided by the Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, The Netherlands.

Author contributions

Stan R.W. Wijn: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - Original Draft, Visualization.

Maroeska M. Rovers: Conceptualization, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

Gerjon Hannink: Conceptualization, Methodology, Validation, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

References

1. Blonde L, Khunti K, Harris SB, Meizinger C, Skolnik NS. Interpretation and Impact of Real-World Clinical Data for the Practicing Clinician. Adv. Ther. 2018;35(11):1763–74.
2. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav. Res. 2011;46(3):399–424.
3. Pazzagli L, Linder M, Zhang M, Vago E, Stang P, Myers D, et al. Methods for time-varying exposure related problems in pharmacoepidemiology: An overview. Pharmacoepidemiol. Drug Saf. 2018;27(2):148–60.
4. Zhang Z, Li X, Wu X, Qiu H, Shi H. Propensity score analysis for time-dependent exposure. Ann. Transl. Med. 2020;8(5):246–246.
5. Lu B. Propensity score matching with time-dependent covariates. Biometrics. 2005;61(3):721–8.
6. Austin PC, Stuart EA. The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. Stat. Methods Med. Res. 2017;26(4):1654–70.
7. Morgan SL, Winship C. Counterfactuals and Causal Inference. Counterfactuals Causal Inference Methods Princ. Soc. Res. Cambridge: Cambridge University Press; 2014. 1–499 p.
8. Robins JM, Hernán MÁ, Brumback B. Marginal Structural Models and Causal Inference in Epidemiology. Epidemiology. 2000;11(5):550–60.
9. Grant MJ, Booth A. A typology of reviews: An analysis of 14 review types and associated methodologies. Health Info. Libr. J. 2009;26(2):91–108.
10. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann. Intern. Med. 2018;169(7):467.
11. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst. Rev. 2016;5(1):210.
12. Sievert C. Interactive Web-Based Data Visualization with R, plotly, and shiny [Internet].

Interact. Web-Based Data Vis. with R, plotly, shiny. Chapman and Hall/CRC; 2020.

13. Clare PJ, Dobbins TA, Mattick RP. Causal models adjusting for time-varying confounding—a systematic review of the literature. Int. J. Epidemiol. 2019;48(1):254–65.

14. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat. Med. 2015;34(28):3661–79.

15. Naimi AI, Cole SR, Kennedy EH. An introduction to g methods. Int. J. Epidemiol. 2017;46(2):756–62.

16. Kupzyk KA, Beal SJ. Advanced Issues in Propensity Scores. J. Early Adolesc. 2017;37(1):59–84.

17. Mansournia MA, Etminan M, Danaei G, Kaufman JS, Collins G. Handling time varying confounding in observational research. BMJ. 2017;359.

18. Dukes O, Vansteelandt S. A Note on G-Estimation of Causal Risk Ratios. Am. J. Epidemiol. 2018;187(5):1079–84.

19. Wijn SRW, Rovers MM, van Tienen TG, Hannink G. Intra-articular corticosteroid injections increase the risk of requiring knee arthroplasty. Bone Joint J. 2020;102-B(5):586–92.

20. Rongen JJ, Rovers MM, van Tienen TG, Buma P, Hannink G. Increased risk for knee replacement surgery after arthroscopic surgery for degenerative meniscal tears: a multi-center longitudinal observational study using data from the osteoarthritis initiative. Osteoarthr. Cartil. 2017;25(1):23–9.

21. Austin PC. A comparison of 12 algorithms for matching on the propensity score. Stat. Med. 2014;33(6):1057–69.

22. Lin V, McGrath S, Zhang Z, Petito LC, Logan RW, Hernán MA, et al. gfoRmula: An R package for estimating effects of general time-varying treatment interventions via the parametric g-formula. 2019;

23. R Core Team. R: A Language and Environment for Statistical Computing. J. Stat. Softw. Vienna, Austria: R Foundation for Statistical Computing; 2017.

24. Therneau TM. A Package for Survival Analysis in S. 2015.

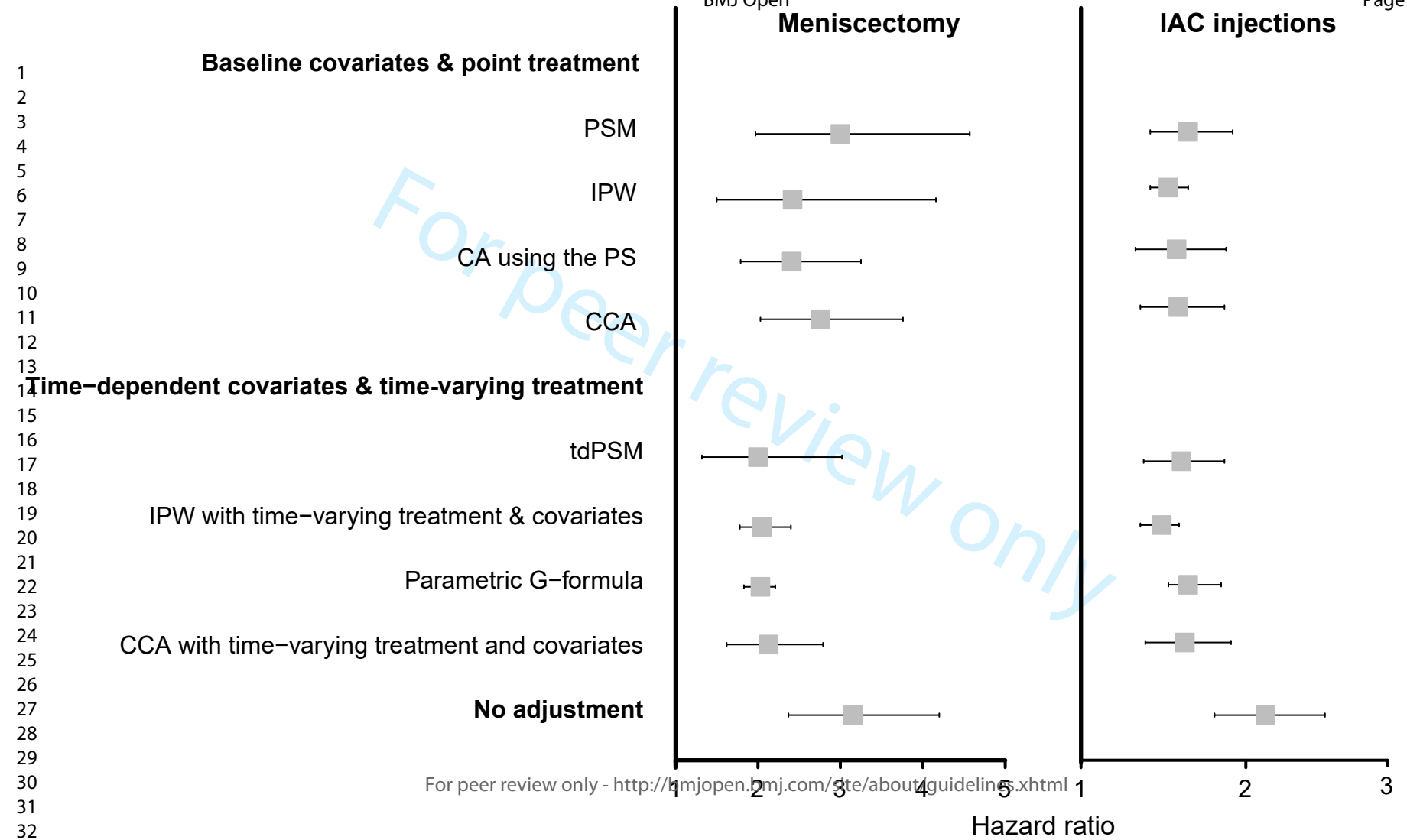
25. Ho DE, Imai K, King G, Stuart EA. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. J. Stat. Softw. 2011;42(8):1–28.

26. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J. Stat. Softw. 2011;45(3):1–67.

27. Greifer N. WeightIt: Weighting for Covariate Balance in Observational Studies. 2019.

28. Dunkler D, Ploner M, Schemper M, Heinze G. Weighted cox regression using the R package coxphw. J. Stat. Softw. 2018;84(2).

29. Canty A, Ripley B. boot: Bootstrap R (S-Plus) Functions. R package version 1.3-15. 2015.



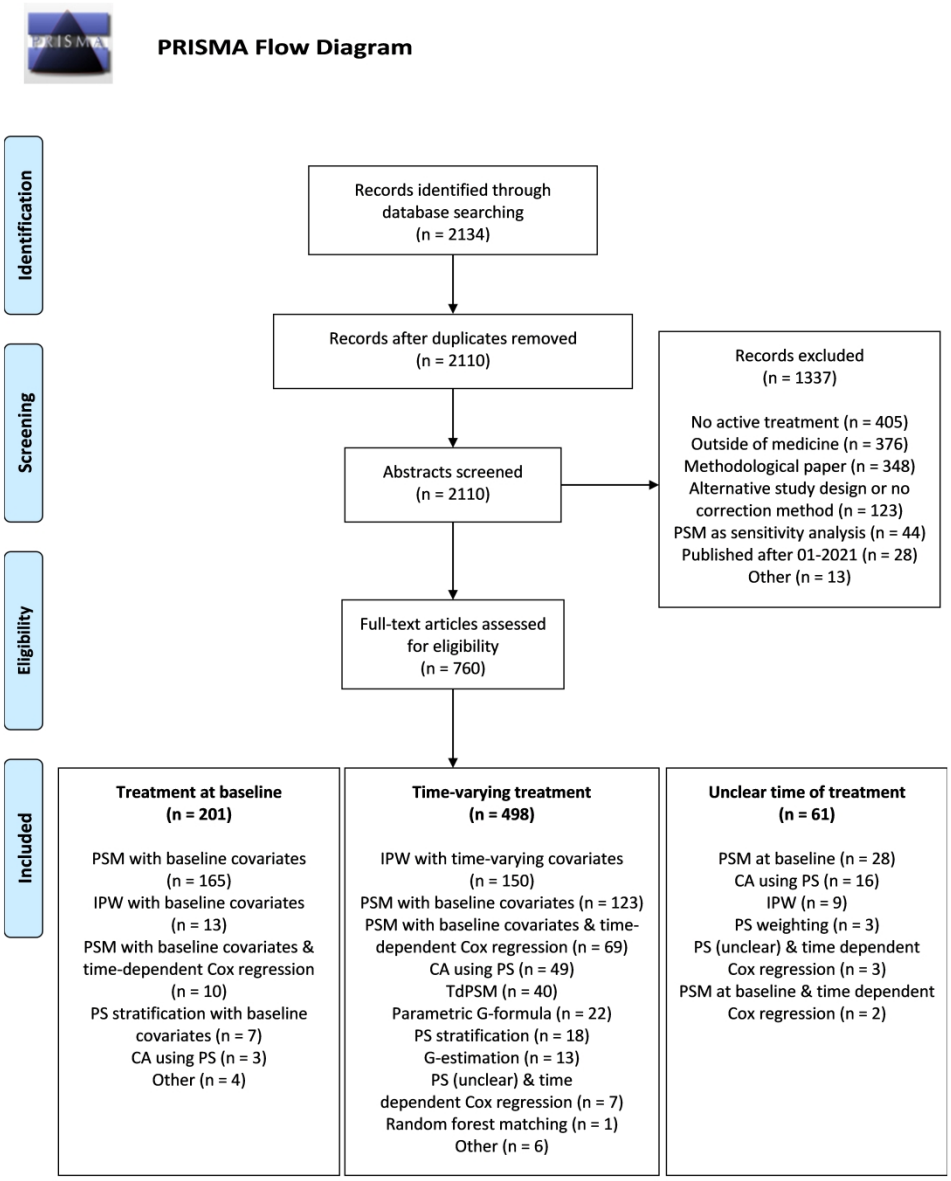
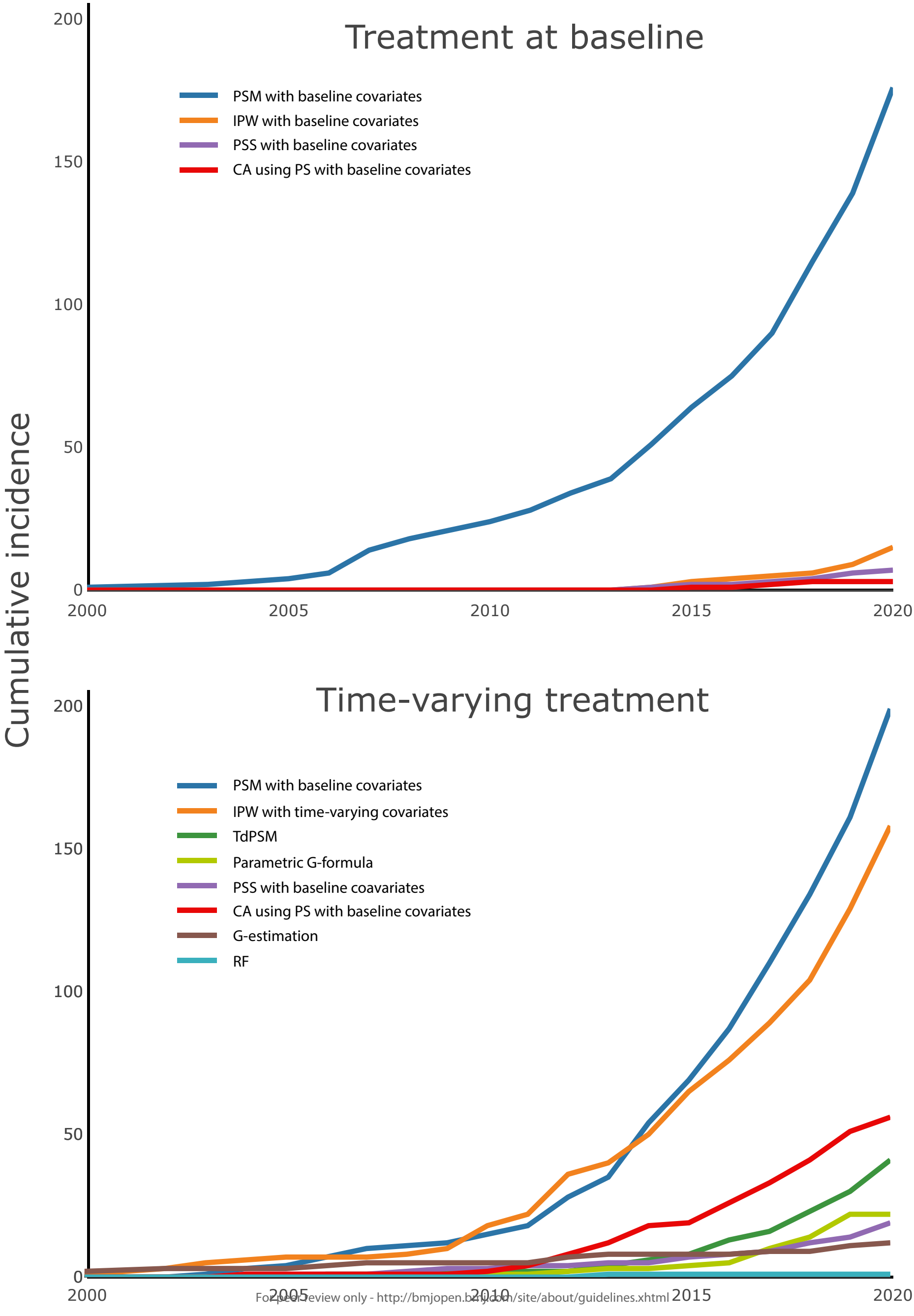


Figure 2: PRISMA Flow Diagram of the flow of papers in the mapping review. In total, 760 studies were included and categorized according to the time of treatment. PSM, propensity score matching; IPW, inverse probability weighting; CA, covariate adjustment; PS, propensity score; TdPSM, time-dependent propensity score matching.

1314x1606mm (96 x 96 DPI)



Supplement S1: Search strategy

Initial search	(time var*[tiab] OR time dependent*[tiab] AND iptw[tiab] OR inverse probability[tiab]) OR ("Propensity Score"[Mesh] OR propensity score*[tiab])	28373
Included propensity mesh to reduce the number of papers	("Propensity Score"[Mesh] OR propensity score*[tiab]) AND (time var*[tiab] OR time dependent*[tiab] OR iptw[tiab] OR inverse probability[tiab])	1566
Added risk-set matching	(risk-set matching[tiab]) OR ("Propensity Score"[Mesh] OR propensity score*[tiab]) AND (time var*[tiab] OR time dependent*[tiab] OR iptw[tiab] OR inverse probability[tiab])	1570
Added g-methods to the search and studies published before 2021	(risk-set matching[tiab]) OR ("Propensity Score"[Mesh] OR propensity score*[tiab]) OR (“g-methods”[tiab] OR “g-formula”[tiab] OR “g-estimation”[tiab] OR “parametric g-formula”[tiab]) OR (iptw[tiab] OR inverse probability[tiab]) AND (time var*[tiab] OR time dependent*[tiab] or longitudinal*[tiab])	2134

Supplement S2: Empirical example details from Box 1

Empirical examples

Data from the Osteoarthritis Initiative (OAI) was used for two empirical examples. The OAI is a multicentre, longitudinal cohort study that included patients with (or at risk for) symptomatic femoral-tibial knee osteoarthritis (OA) with a follow-up up to 108 months, available for public access at https://data-archive.nimh.nih.gov/oai/. We extracted a large set of variables from the OAI that were measured at baseline and annual follow-up visits (12 to 108 months), these include general patients characteristics (age, gender, history of knee symptoms, physical activity, weight, care access), clinical variables (knee symptoms, radiographic signs of OA, hand OA), quality of life measurements (12-Item Short Form Survey (SF-12)), functional scores (Knee injury and Osteoarthritis Outcome Score (KOOS), Western Ontario and McMasters Osteoarthritis index (WOMAC)) and time-varying treatments (meniscectomy, knee replacement surgery, corticosteroid injections). Missing values were imputed through single imputation using predictive mean matching for continuous variables and logistic regression for categorical variables.

To investigate the impact of the different confounding adjustment methods on the outcome, two empirical examples with a time-varying treatment were selected that we previously published using data from the OAI: 1) the effect of meniscectomy (surgical removal of the meniscus) on the risk to receive knee replacement surgery and 2) the effect of intra-articular corticosteroid injections on the risk to receive knee replacement surgery.[19,20]

Statistical methods

In total, we compared nine methods that were the most commonly used adjustment methods found in the mapping review for both empirical examples: four methods that matched using baseline covariates, four time-dependent methods, and no matching. Confounding factors included in all eight correction methods were: patient characteristics (age, gender, BMI, physical activity, health care access, treatment centre, education, family history with OA, occupation), clinical variables (knee

medication use, hand OA at baseline, knee symptoms at baseline, radiographic confirmed OA), quality of life scores (SF-12 subscales), and functional scores (KOOS and WOMAC). After adjustment, Cox proportional hazard models were applied to estimate the treatment effect and confidence intervals.

The baseline methods consisted of PSM, IPW with a point treatment (yes/no), covariate adjustment using the propensity score, and conventional covariate adjustment (CCA) using baseline covariates and a point treatment. For PSM, the propensity score was calculated for every patient (the probability of a patient being assigned to the treatment given a set of observed covariates) and subsequently treated and control patients were matched using a 1:1 matching ratio without replacement, a caliper of 0.20 and a nearest neighbour matching algorithm, as nearest neighbour is commonly used and results in less biased estimates compared to the other matching algorithms.[21] Covariate balance was assessed by calculating the standardized mean difference (SMD) and by plotting the balance between patients and controls. Balance smaller or equal to 0.10 SMD were assumed to have appropriate balance.[2] IPW was performed to build a marginal structural model able to balance the covariates at baseline (marginal structural model with point treatment; patients were either labelled as treated or untreated). For IPW we used unbalanced weights and the weights were visually inspected. Similar to PSM, a 0.10 SMD was assumed to have an appropriate balance. Confidence intervals were estimated using 1000 bootstraps. Covariate adjustment using the propensity score was performed by calculating the propensity score using logistic regression and subsequently the propensity score was added to the Cox regression. Conventional covariate adjustment was performed by including the same set of covariates in the Cox regression without any prior adjustment.

The time-dependent methods consisted of time-dependent propensity score matching (tdPSM), IPW with time-varying treatment, parametric g-formula, and CCA with time-varying treatment and covariates.[5,15,17] Time-dependent propensity score matching was performed by sequentially

matching treated patients with all available controls at time of treatment using a 1:1 nearest neighbour matching algorithm without replacement using a caliper of 0.2. After matching a patient to a control, both were removed from the dataset to avoid further matches. Similar to the baseline methods, IPW was used to create a marginal structural model but with time-varying treatment and time-varying covariates. Likewise, we used unbalanced weights and the weights were visually inspected and balance was assessed. Confidence intervals were estimated using 1000 bootstraps.

Robins' g-formula (also known as parametric g-formula or parametric g-computation) is an alternative method to recover effects of time-varying treatment under untestable assumptions, given that sufficient covariates are measured to control for confounding by unmeasured risk factors.[22] The causal effect is measured by comparing the treatment effect between an always exposed- and a never exposed scenario. Conventional covariate adjustment with time-varying covariates and treatment was performed by including these variables in the Cox regression.

Finally, we performed one crude analysis by only including the time-varying treatment in the Cox regression. All analyses and simulations were performed using R (version 4.0.2, The R Foundation for Statistical Computing, Vienna, Austria) using packages 'mice', 'MatchIt', 'WeightIt', 'gfoRmula', 'plotly', 'coxphw', 'boot', and 'survival'.[12,22–29]

Results

In total, nine methods were compared for both empirical examples: four methods that adjust using baseline covariates (PSM, IPW using point treatment, CA using the propensity score, CCA), four time-dependent methods (tdPSM, IPW using time-varying treatment, parametric g-formula, CCA) and one without adjustments. (see figure in Box 1)

In the meniscectomy example, patients who underwent meniscectomy had an HR of 3.0 (95% CI: 1.97– 4.57), 2.42 (95% CI: 1.50 – 4.16), 2.41 (95% CI: 1.79 – 3.25), and 2.76 (95% CI: 2.03 – 3.76) to receive knee replacement surgery for PSM, IPW, CA using the propensity score, and CCA using the baseline covariates, respectively. The time-dependent strategies resulted in lower hazard ratios: HR of 2.00 (95% CI: 1.32 – 3.02), 2.05 (95% CI: 1.78 – 2.40), 2.03 (95% CI: 1.83 – 2.21) and 2.13 (95% CI: 1.62 – 2.79) for tdPSM, IPW, parametric G-formula and CCA, respectively. Without any adjustment, an HR of 3.15 (95% CI: 2.37 – 4.20) was found.

The results from intra-articular corticosteroid injection examples were more consistent between the baseline and time-dependent methods. Patients that receive intra-articular corticosteroid injections had a higher risk to receive knee replacement surgery with an HR of 1.64 (95% CI: 1.42 – 1.92), 1.53 (95% CI: 1.42 – 1.65), 1.58 (95% CI: 1.33 – 1.88), and 1.59 (95% CI: 1.36 – 1.87) for the baseline methods (PSM, IPW, CA using the propensity score, and CCA, respectively) and an HR of 1.61 (95% CI: 1.38 – 1.87), 1.49 (95% CI: 1.36 – 1.57), 1.65 (95% CI: 1.53 – 1.85) and 1.63 (95% CI: 1.39 – 1.91) for the time-dependent methods (tdPSM, IPW, parametric g-formula, CCA, respectively). No adjustment resulted in an HR of 2.12 (95% CI: 1.81 – 2.48).

Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
TITLE			
Title	1	Identify the report as a scoping review.	1
ABSTRACT			
Structured summary	2	Provide a structured summary that includes (as applicable): background, objectives, eligibility criteria, sources of evidence, charting methods, results, and conclusions that relate to the review questions and objectives.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known. Explain why the review questions/objectives lend themselves to a scoping review approach.	4
Objectives	4	Provide an explicit statement of the questions and objectives being addressed with reference to their key elements (e.g., population or participants, concepts, and context) or other relevant key elements used to conceptualize the review questions and/or objectives.	4
METHODS			
Protocol and registration	5	Indicate whether a review protocol exists; state if and where it can be accessed (e.g., a Web address); and if available, provide registration information, including the registration number.	-
Eligibility criteria	6	Specify characteristics of the sources of evidence used as eligibility criteria (e.g., years considered, language, and publication status), and provide a rationale.	5
Information sources*	7	Describe all information sources in the search (e.g., databases with dates of coverage and contact with authors to identify additional sources), as well as the date the most recent search was executed.	5
Search	8	Present the full electronic search strategy for at least 1 database, including any limits used, such that it could be repeated.	5 & Supplement S1
Selection of sources of evidence†	9	State the process for selecting sources of evidence (i.e., screening and eligibility) included in the scoping review.	5
Data charting process‡	10	Describe the methods of charting data from the included sources of evidence (e.g., calibrated forms or forms that have been tested by the team before their use, and whether data charting was done independently or in duplicate) and any processes for obtaining and confirming data from investigators.	5
Data items	11	List and define all variables for which data were sought and any assumptions and simplifications made.	5
Critical appraisal of individual sources of evidence§	12	If done, provide a rationale for conducting a critical appraisal of included sources of evidence; describe the methods used and how this information was used in any data synthesis (if appropriate).	5

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
Synthesis of results	13	Describe the methods of handling and summarizing the data that were charted.	6
RESULTS			
Selection of sources of evidence	14	Give numbers of sources of evidence screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally using a flow diagram.	8 & 9
Characteristics of sources of evidence	15	For each source of evidence, present characteristics for which data were charted and provide the citations.	8
Critical appraisal within sources of evidence	16	If done, present data on critical appraisal of included sources of evidence (see item 12).	-
Results of individual sources of evidence	17	For each included source of evidence, present the relevant data that were charted that relate to the review questions and objectives.	8
Synthesis of results	18	Summarize and/or present the charting results as they relate to the review questions and objectives.	8
DISCUSSION			
Summary of evidence	19	Summarize the main results (including an overview of concepts, themes, and types of evidence available), link to the review questions and objectives, and consider the relevance to key groups.	10
Limitations	20	Discuss the limitations of the scoping review process.	10
Conclusions	21	Provide a general interpretation of the results with respect to the review questions and objectives, as well as potential implications and/or next steps.	12
FUNDING			
Funding	22	Describe sources of funding for the included sources of evidence, as well as sources of funding for the scoping review. Describe the role of the funders of the scoping review.	13

JBI = Joanna Briggs Institute; PRISMA-ScR = Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews.
* Where *sources of evidence* (see second footnote) are compiled from, such as bibliographic databases, social media platforms, and Web sites.
† A more inclusive/heterogeneous term used to account for the different types of evidence or data sources (e.g., quantitative and/or qualitative research, expert opinion, and policy documents) that may be eligible in a scoping review as opposed to only studies. This is not to be confused with *information sources* (see first footnote).
‡ The frameworks by Arksey and O'Malley (6) and Levac and colleagues (7) and the JBI guidance (4, 5) refer to the process of data extraction in a scoping review as data charting.
§ The process of systematically examining research evidence to assess its validity, results, and relevance before using it to inform a decision. This term is used for items 12 and 19 instead of "risk of bias" (which is more applicable to systematic reviews of interventions) to include and acknowledge the various sources of evidence that may be used in a scoping review (e.g., quantitative and/or qualitative research, expert opinion, and policy document).

From: Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMAScR): Checklist and Explanation. Ann Intern Med. 2018;169:467–473. doi: 10.7326/M18-0850.

BMJ Open

Confounding adjustment methods in longitudinal observational data with a time-varying treatment: a mapping review

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-058977.R1
Article Type:	Original research
Date Submitted by the Author:	02-Feb-2022
Complete List of Authors:	Wijn, Stan; Radboudumc, Radboud university medical center, Radboud Institute for Health Sciences, Department of Operating Rooms Rovers, Maroeska; Radboudumc, Radboud university medical center, Radboud Institute for Health Sciences, Department of Operating Rooms and Health Evidence Hannink, Gerjon; Radboudumc, Radboud university medical center, Radboud Institute for Health Sciences, Department of Operating Rooms
Primary Subject Heading:	Research methods
Secondary Subject Heading:	Epidemiology, Evidence based practice
Keywords:	EPIDEMIOLOGY, STATISTICS & RESEARCH METHODS, Orthopaedic & trauma surgery < SURGERY

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Confounding adjustment methods in longitudinal observational data with a time-varying treatment: a mapping review

Stan R.W. Wijn[1], Maroeska M. Rovers[1,2], Gerjon Hannink[1]

[1] Radboud University Medical Centre, Radboud Institute for Health Sciences, Department of Operating Rooms, Nijmegen, the Netherlands

[2] Radboud University Medical Centre, Radboud Institute for Health Sciences, Department of Health Evidence, Nijmegen, the Netherlands

S.R.W. Wijn, Stan.Wijn@radboudumc.nl

M.M. Rovers, Maroeska.Rovers@radboudumc.nl

G. Hannink, Gerjon.Hannink@radboudumc.nl

Corresponding author:

Stan R.W. Wijn
Radboud university medical centre
715 Department of Operating Rooms
P.O. Box 9101
6500 HB Nijmegen
The Netherlands

Declarations of interest: none

Word count: 2114

Keywords: Propensity score matching, longitudinal observational data, time-varying treatment, confounding, g-methods

Abstract

Objectives

To adjust for confounding in observational data, researchers use propensity score matching (PSM), but more advanced methods might be required when dealing with longitudinal data and time-varying treatments as PSM might not include possible changes that occurred over time. This study aims to explore which confounding adjustment methods have been used in longitudinal observational data to estimate a treatment effect and identify potential inappropriate use of PSM.

Design

Mapping review.

Data sources

We searched PubMed, from inception up to January 2021, for studies in which a treatment was evaluated using longitudinal observational data.

Eligibility criteria

Methodological-, non-medical- and cost-effectiveness papers were excluded, as were non-English studies and studies that did not study a treatment effect.

Data extraction and synthesis

Studies were categorized based on time of treatment: at baseline (interventions performed at start of follow-up) or time-varying (interventions received asynchronously during follow-up) and sorted based on publication year, time of treatment and confounding adjustment method. Cumulative time series plots were used to investigate the use of different methods over time. No risk-of-bias assessment was performed as it was not applicable.

Results

In total, 764 studies were included that met the eligibility criteria. PSM (165/201, 82%) and inverse probability weighting (154/502, 31%) were most common for studies with a treatment at baseline (n=201) and time-varying treatment (n=502), respectively. Of the 502 studies with a time-varying treatment, 123 (25%) used PSM with baseline covariates, which might be inappropriate. In the past five years, the proportion of studies with a time-varying treatment that used PSM over inverse probability weighting increased.

Conclusions

PSM is the most frequently used method to correct for confounding in longitudinal observational data. In studies with a time-varying treatment, PSM was potentially inappropriately used in 25% of studies. Confounding adjustment methods designed to deal with a time-varying treatment and time-varying confounding are available, but were only used in 45% of the studies with a time-varying treatment.

Strengths and limitations of this study

- We systematically mapped the literature from inception up to January 2021 for the most commonly used methods to correct for confounding in longitudinal observational data.
- This study was conducted and reported according to the PRISMA extension for scoping reviews (PRISMA-ScR)
- No risk-of-bias assessment was performed because the scope of this mapping review targets the statistical methods that have been used in the included studies, so a risk of bias assessment was not applicable.
- For some studies we were not able to identify if patients were treated at baseline or during follow-up (fortunately, this issue was only apparent in 8% of the included studies).

Introduction

The increasing availability of real-world data derived from electronic health records, registries, wearables, and surveys can be a valuable source of data to evaluate the effectiveness of a treatment.[1] Deriving inference directly from real-world data can be challenging as it is prone to confounding. To adjust for confounding, researchers use methods such as propensity score matching (PSM) to create two comparable groups in which both the treated- and untreated patients have similar observable characteristics (like age, pain scores, weight etc.) similar to a randomised trial.[2] Although these methods can be sufficient when a patient is treated at the start of a study (baseline), more advanced methods might be required when dealing with longitudinal data and time-varying or repeated treatments. Adjustment at baseline in the presence of longitudinal data and time-varying treatment might not include possible changes that occurred over time. These can include changes in treatment regimens or disease progression, but can also comprise weight changes, pain scores or changes in behaviour (e.g., stopped smoking). These changes can alter the balance between treated- and untreated patients and can result in different estimates of the treatment effect (see box 1).[3,4] Methods like time-dependent propensity score matching and the g-methods (inverse probability weighting (IPW), parametric g-formula or g-estimation) can incorporate time-varying covariates and time-varying treatments and can take feedback between the treatment and outcome over time into account.[2,5–8] It is however unclear if these methods are regularly used in practice when dealing with longitudinal observational data with a time-varying treatment. Therefore, this mapping review aimed to identify and describe which methods have been used to adjust for confounding bias in longitudinal observational data and identify potential inappropriate use of baseline adjustment methods (like PSM).

Box 1: Empirical example using data from the Osteoarthritis Initiative

To investigate the influence of the different confounding adjustment methods on the outcome, two previously published empirical examples with a time-varying treatment were selected: 1) the effect of meniscectomy (surgical removal of the meniscus) and 2) the effect of intra-articular corticosteroid injections on the risk to receive knee replacement surgery.[9,10] Data from the Osteoarthritis Initiative (OAI) was used for both examples. The OAI is a multicentre, longitudinal cohort study that included patients with (or at risk for) symptomatic femoral-tibial knee osteoarthritis (OA) with a follow-up up to 108 months, available for public access at https://data-archive.nimh.nih.gov/oai/. A large set of variables was extracted from the OAI, measured at baseline and annual follow-up visits. These include general patient characteristics, clinical variables, quality of life measurements, functional scores and time-varying treatments.

In total, we compared nine commonly used adjustment for both empirical examples: four methods that corrected using baseline covariates, four time-dependent methods, and no matching. We found in the first example (meniscectomy) that adjustment using baseline covariates resulted in larger estimates of the treatment effect compared to time-dependent methods, while results were consistent in the second example (intra-articular corticosteroid injection; figure 1). These results show that the selected adjustment method can influence the detected treatment effect when dealing with potential time-varying confounding. See Supplement S2 for more details.

<insert figure 1>

Figure 1: Forest plot displaying the results of the two empirical examples (left: meniscectomy, right: intra-articular corticosteroid (IAC)). Four methods were compared using baseline covariates, four methods using time-dependent covariates and time-varying treatment and one without correction. PSM, propensity score matching; IPW, inverse probability weighting; CCA, conventional covariate adjustment; IAC, intra-articular corticosteroids; tdPSM, time-dependent propensity score matching.

Methods

A mapping literature review was performed to determine which confounding adjustment methods were used in longitudinal observational data to estimate a treatment effect. Mapping reviews are designed to map out and categorize existing literature and explore trends and identify gaps by study design and other key features.[11] This study was conducted and reported according to The PRISMA extension for Scoping Reviews (PRISMA-ScR).[12]

Patient and public involvement

Patients and/or public were not involved.

Search strategy

We searched in PubMed from inception up to January 2021 for papers in which a treatment was evaluated using longitudinal observational data. Search terms used were time varying, longitudinal observational data, and commonly used adjustment methods and terms (e.g., matching, g-methods). The search strategy can be found in Supplement SI. Methodological-, non-medical- and cost-effectiveness papers were excluded as well as non-English studies or studies that did not study a treatment effect. Studies that used no adjustment method or used the adjustment method solely as sensitivity analysis were also excluded.

All papers were screened based on title and abstract and papers that met the inclusion criteria were screened full-text. The title, author(s), journal, research theme, publication date, confounding adjustment method, and time of treatment (at baseline or time-varying) were extracted from all papers that met the inclusion criteria. A treatment at baseline was defined as an intervention performed at the start of follow-up for all included patients (e.g., all treated patients received surgery at the start of follow-up). Time-varying treatment was defined as a treatment that was received asynchronously during follow-up (e.g., patients received surgery at different moments during follow-up) or when dealing with a repeated treatment of which the timing was not identical

for all treated patients (e.g., personalized medication intake over time). If the time of treatment was not defined, studies were categorized as unclear.

Study selection and data extraction was performed by one reviewer (SW). Any issues during study selection, data extraction or analysis were discussed and resolved by all authors. No risk of bias assessment was performed because the scope of this paper targets the statistical methods that have been used in these papers, and therefore a risk of bias assessment was not applicable.

Analysis

Study selection was performed in Rayyan.[13] Study characteristics (author, publication year, journal), time of treatment (at baseline, time-varying or unclear) and confounding adjustment method were extracted and analysed in R (version 4.1.0, The R Foundation for Statistical Computing, Vienna, Austria). Studies were sorted based on publication year, time of treatment and confounding adjustment method and described using descriptive statistics. If a study used multiple adjustment methods or a combination of methods, we included all methods, i.e., more methods than papers could be identified. Cumulative time series plots were used to investigate the use of different methods over time for treatments at baseline and time-varying treatments using the Plotly package.[14]

Results

Our search identified 2140 articles of which eventually 764 met the eligibility criteria after title and abstract review, and subsequent full-text review (see also figure 2). The main reasons for exclusion were the lack of intervention/treatment (n = 405), a scope outside of medicine (n = 376), a methodological paper (n = 348), or the study did not utilize longitudinal observational data or did not correct for confounding (n = 123). Of all included papers, 201 (26%) had a treatment at baseline, 502 (66%) had a time-varying treatment and 61 (8%) papers had no clearly defined time of treatment. Of the papers with a treatment at baseline, the majority used PSM with baseline covariates (n = 165, 82%) as a method to correct for confounding. Studies that had a time-varying treatment most often used IPW (154 papers, 30%), PSM with baseline covariates was used in 123 papers (25%), PSM with baseline covariates combined with time-dependent Cox regression in 69 papers (14%), covariate adjustment using the propensity score in 49 papers (10%), time-dependent PSM in 40 papers (8%), parametric G-formula in 22 papers (4%), propensity score stratification in 18 papers (2%) and G-estimation in 13 papers (3%). Confounding adjustment methods designed to deal with a time-varying treatment and time-varying confounding (IPW, parametric g-formula or g-estimation) were used in 45% of the papers with a time-varying treatment. In the last five years, the proportion of studies with a time-varying treatment that used PSM with baseline covariates over IPW increased (199 vs 158 in 2020, for PSM with baseline covariates and IPW, respectively). (Figure 3) For papers of which the time of treatment was unclear, PSM at baseline was most frequently used in 28 papers (46%). We added an overview of the most commonly used methods found in our search and when they should be used. (Figure 4)

<insert Figure 2>

Figure 2: PRISMA Flow Diagram of the flow of papers in the mapping review. In total, 764 studies were included and categorized according to the time of treatment. PSM, propensity score matching; IPW, inverse probability weighting; CA, covariate adjustment; PS, propensity score; TdPSM, time-dependent propensity score matching.

<insert Figure 3>

Figure 3: Cumulative incidence of the different confounding adjustment methods that are used in practice. Some studies used multiple methods. PSM, propensity score matching; IPW, inverse probability weighting; CA, covariate adjustment; PS, propensity score; TdPSM, time-dependent propensity score matching; PSS, propensity score stratification; RF, random forest matching.

<insert Figure 4>

Figure 4: Common methods to correct for confounding and when they should be used.

Discussion

Although advanced methods are available to correct for confounding in longitudinal observational data, we showed that these methods are not always utilized in studies that have a time-varying treatment. Instead, 25% of the studies that had a time-varying treatment used PSM with baseline covariates to correct for confounding which can potentially result in a biased treatment effect.[4]

Our findings confirm the results by Clair et al. whom provided a summary of new methods that have been used in literature to deal with time-varying confounding. They concluded that IPW was most commonly used, more robust methods (like g-estimation) were underused.[15] Our results are also in agreement with the findings by Austin et al. whom reported a rapidly increasing use of IPW in the literature in the last decade.[16] Nonetheless, we detected a similarly rapid growth in the use of PSM in studies with a time-varying treatment, which can potentially result in biased results as PSM does not correct for time-varying confounding. Although time-dependent methods like tdPSM, parametric g-formula and IPW are extensively described in the literature [5,8,17], adjusting at baseline in observational data is still common in literature and was used in 25% of the papers with a time-varying treatment we included in our mapping review.[18] The proportion of studies with a time-varying treatment that used PSM over IPW even increased in the last five years.

Some potential limitations should also be discussed. First, the main limitation of a mapping review is the broad descriptive level at which studies are analysed and described. However, it does provide a general overview of the published literature and suggests that methods to deal with confounding in studies with a time-varying treatment are underused. Furthermore, no risk of bias assessment of the included studies was performed and study selection and data extraction were performed by one reviewer. Using a second reviewer throughout the entire study screening process could increase the number of relevant studies identified for use in a systematic review.[19] However, as we targeted the overall trends in data analysis of studies with longitudinal observational data, this would likely not affect our conclusions much. Second, although it is common to search multiple databases in a

systematic review, our mapping review was limited to PubMed. We found over 2000 papers in PubMed which was ample for the aim of this study and for a mapping review. It is unlikely that additional searches could alter our conclusions. Third, for some studies we were not able to identify if patients were treated at baseline or during follow-up. Fortunately, this only occurred in 8% of the papers we included.

Implications

From previously published studies we can conclude that time-dependent methods can be important to avoid biased estimates of the treatment effect when adjusting for confounding in longitudinal observational data with potential time-varying confounding.[4,20] Therefore, we suggest using one of the g-methods (IPW, parametric g-formula, g-estimation) with time-varying covariates and time-varying treatment if the data is available.[20] Yet, these methods are not the panacea for unconfounded analyses in longitudinal observational data. They still rely on relevant confounder selection (based on prior knowledge, possibly supported by a directed acyclic graph), require careful examination of weights and adequate covariate balance.[16] Although there are clear benefits and limitations to each g-method, it is often unclear what the most appropriate method is to correct for confounding.[17] From the g-methods, IPW has three main advantages over the other methods: 1) it is a commonly used method, 2) it is relatively simple to understand and explain, and 3) it is easy to perform in standard statistical software (like R or STATA). Parametric g-formula is ideal for joint interventions or dynamic interventions but requires more computational power and additional programming.[20] G-estimation is particularly useful for studying the interaction between treatment and time-varying confounders (treatment-confounder feedback), but it can be challenging to implement g-estimation in longitudinal data. G-estimation can also be complex as there are not many practical guidelines or statistical packages that support this method for longitudinal data with a time-varying treatment. The developers of *gesttools* R-package (General Purpose G estimation in R) are currently drafting a comprehensive introduction including an explanation of the structural nested

mean model types, the g-estimation algorithm, instructions to set up the users' dataset, and a tutorial to perform g-estimation.[21]

When dealing with real-world data, g-methods are recommended to evaluate the effectiveness of a treatment to preclude confounding. However, a proper assessment of the required confounding adjustment methods prior to data analysis is appropriate. As we have seen in Box 1, different confounding adjustment methods can potentially influence the conclusions of a study. It depends on many (unknown) case-specific aspects and thus it can be challenging to predict how different methods can affect the conclusion of a study. A direct comparison of different methods to correct for confounding is not recommended as this could stimulate selective reporting of (positive) study results. Every analysis of longitudinal observational data should start by selecting the method best suited for the data at hand. Figure 4 provides an overview of the most commonly used methods and can assist researchers to select the most appropriate method available.

Conclusion

PSM using baseline covariates is the most used method to correct for confounding in longitudinal observational data, even in the presence of a time-varying treatment. Of the 502 identified studies with a time-varying treatment, 123 (25%) used PSM with baseline covariates, which might be inappropriate. Confounding adjustment methods designed to deal with a time-varying treatment and time-varying confounding (IPW, parametric g-formula or g-estimation) are available, but were only used in 45% of the papers with a time-varying treatment and this can potentially result in biased estimates of the treatment effect.

Declarations

Competing interests

The authors declare that they have no competing interest.

Data availability statement

The search strategy is available in the supplemental file and all data extraction documents are available on request to the corresponding author.

Ethics approval statement

Not applicable.

Funding

This work was supported by the Junior Research project (2018) grant provided by the Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, The Netherlands.

Contributors

Stan R.W. Wijn: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - Original Draft, Visualization.

Maroeska M. Rovers: Conceptualization, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

Gerjon Hannink: Conceptualization, Methodology, Validation, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

References

1. Blonde L, Khunti K, Harris SB, Meizinger C, Skolnik NS. Interpretation and Impact of Real-World Clinical Data for the Practicing Clinician. Adv. Ther. 2018;35(11):1763–74.
2. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav. Res. 2011;46(3):399–424.
3. Pazzagli L, Linder M, Zhang M, Vago E, Stang P, Myers D, et al. Methods for time-varying exposure related problems in pharmacoepidemiology: An overview. Pharmacoepidemiol. Drug Saf. 2018;27(2):148–60.
4. Zhang Z, Li X, Wu X, Qiu H, Shi H. Propensity score analysis for time-dependent exposure. Ann. Transl. Med. 2020;8(5):246–246.
5. Lu B. Propensity score matching with time-dependent covariates. Biometrics. 2005;61(3):721–8.
6. Austin PC, Stuart EA. The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. Stat. Methods Med. Res. 2017;26(4):1654–70.
7. Morgan SL, Winship C. Counterfactuals and Causal Inference. Counterfactuals Causal Inference Methods Princ. Soc. Res. Cambridge: Cambridge University Press; 2014. 1–499 p.
8. Robins JM, Hernán MÁ, Brumback B. Marginal Structural Models and Causal Inference in Epidemiology. Epidemiology. 2000;11(5):550–60.
9. Wijn SRW, Rovers MM, van Tienen TG, Hannink G. Intra-articular corticosteroid injections increase the risk of requiring knee arthroplasty. Bone Joint J. 2020;102-B(5):586–92.
10. Rongen JJ, Rovers MM, van Tienen TG, Buma P, Hannink G. Increased risk for knee replacement surgery after arthroscopic surgery for degenerative meniscal tears: a multi-center longitudinal observational study using data from the osteoarthritis initiative. Osteoarthr. Cartil. 2017;25(1):23–9.
11. Grant MJ, Booth A. A typology of reviews: An analysis of 14 review types and associated

methodologies. Health Info. Libr. J. 2009;26(2):91–108.

12. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann. Intern. Med. 2018;169(7):467.

13. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst. Rev. 2016;5(1):210.

14. Sievert C. Interactive Web-Based Data Visualization with R, plotly, and shiny [Internet]. Interact. Web-Based Data Vis. with R, plotly, shiny. Chapman and Hall/CRC; 2020.

15. Clare PJ, Dobbins TA, Mattick RP. Causal models adjusting for time-varying confounding—a systematic review of the literature. Int. J. Epidemiol. 2019;48(1):254–65.

16. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat. Med. 2015;34(28):3661–79.

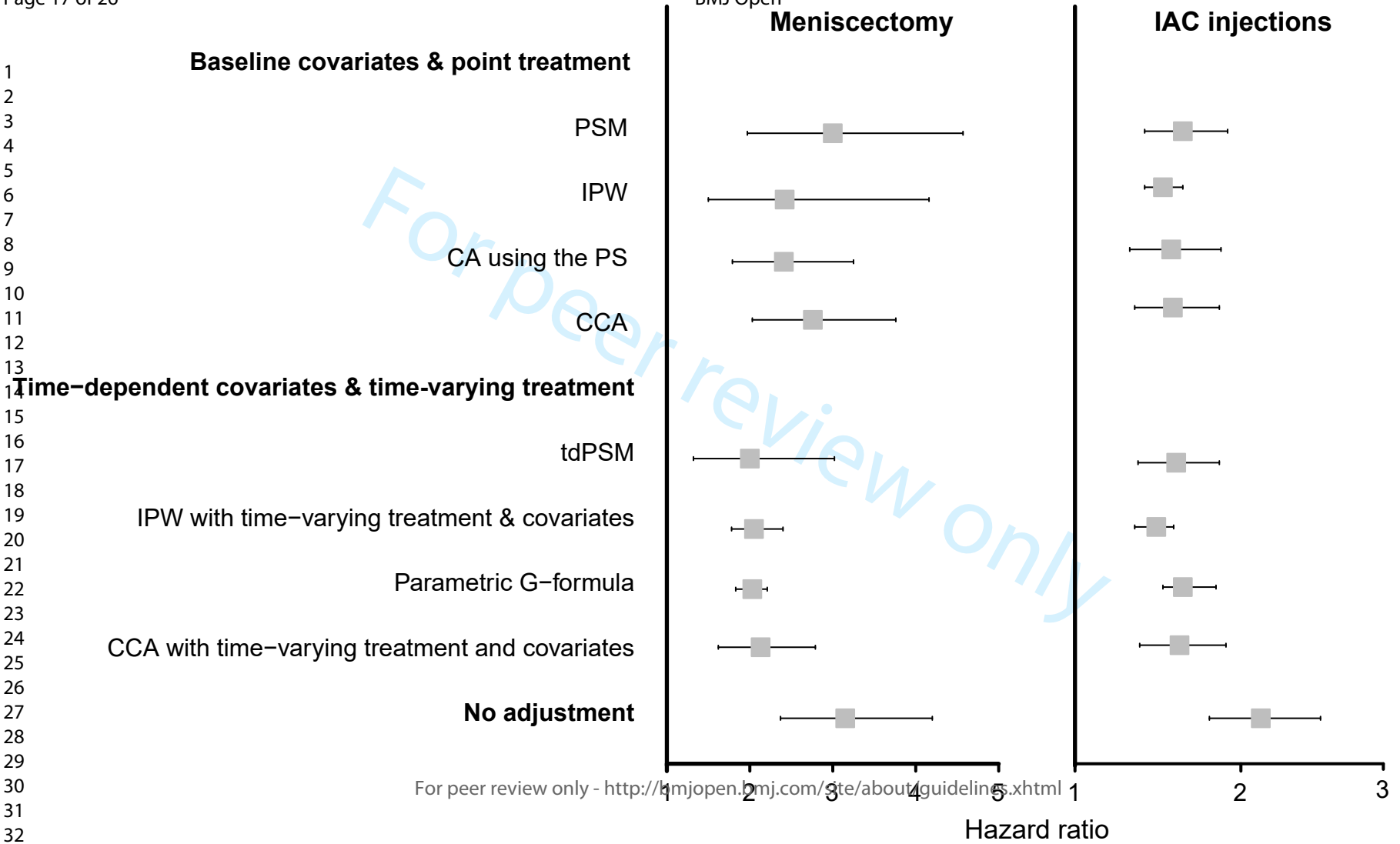
17. Naimi AI, Cole SR, Kennedy EH. An introduction to g methods. Int. J. Epidemiol. 2017;46(2):756–62.

18. Kupzyk KA, Beal SJ. Advanced Issues in Propensity Scores. J. Early Adolesc. 2017;37(1):59–84.

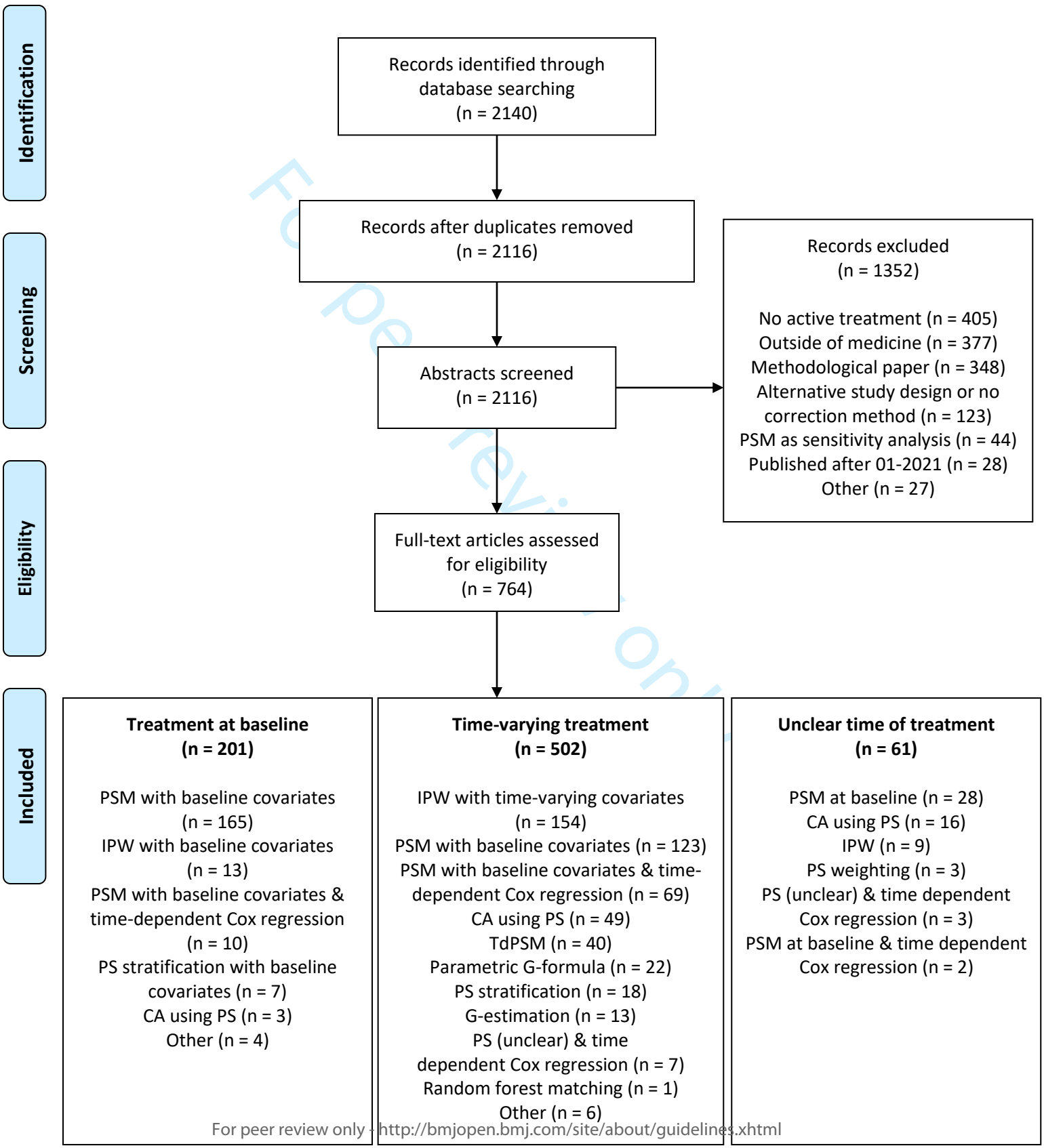
19. Stoll CRT, Izadi S, Fowler S, Green P, Suls J, Colditz GA. The value of a second reviewer for study selection in systematic reviews. Res. Synth. Methods. 2019;10(4):539–45.

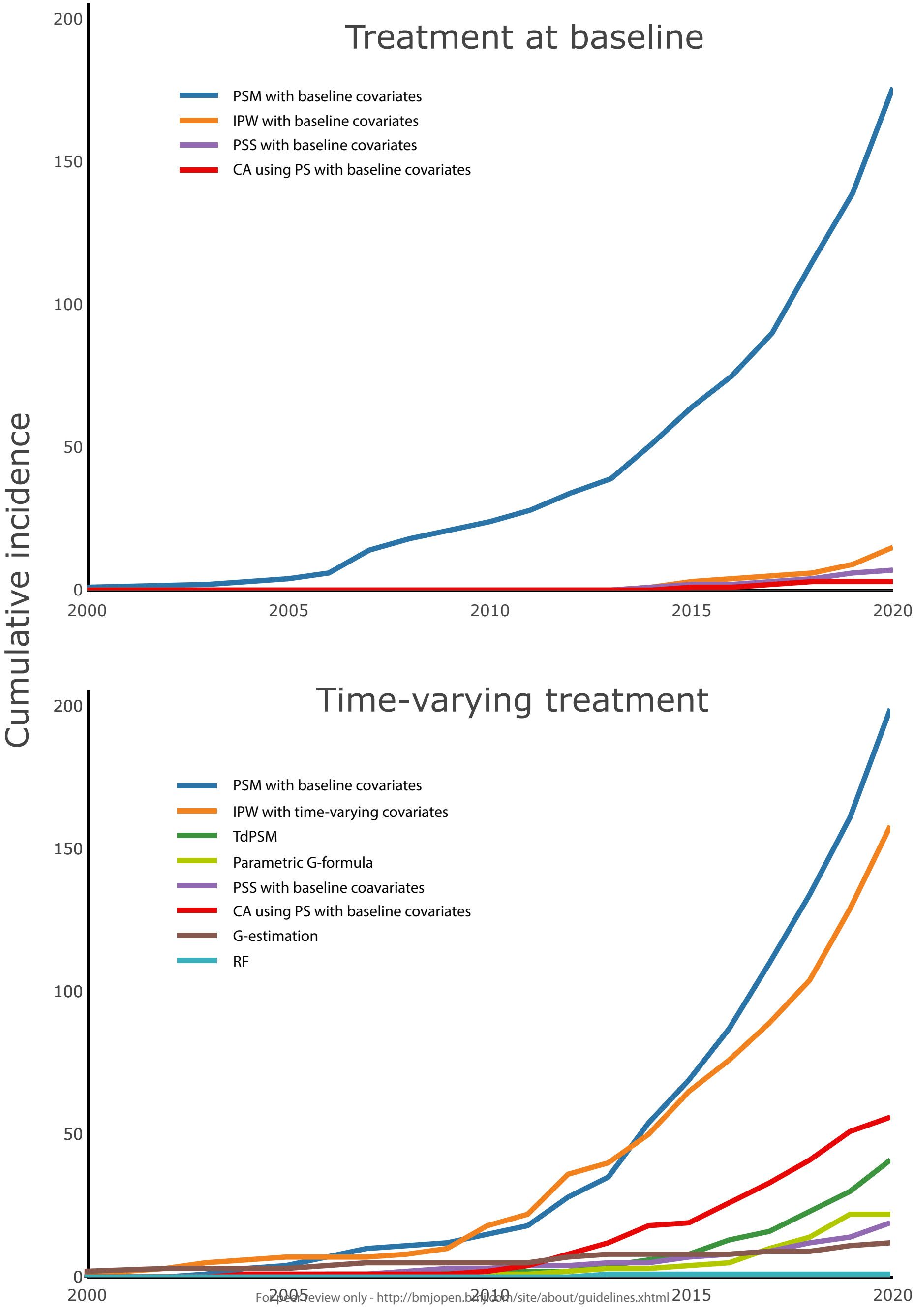
20. Mansournia MA, Etminan M, Danaei G, Kaufman JS, Collins G. Handling time varying confounding in observational research. BMJ. 2017;359.

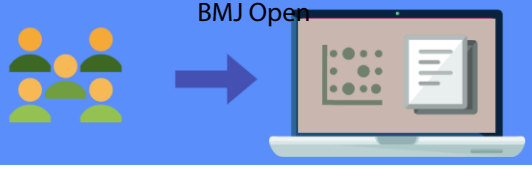
21. Dukes O, Vansteelandt S. A Note on G-Estimation of Causal Risk Ratios. Am. J. Epidemiol. 2018;187(5):1079–84.



PRISMA Flow Diagram







Common methods to correct for confounding

Multiple methods are used to correct for confounding. Here we list the most common types and when they should be used.

Covariate adjustment using propensity score

- The outcome variable is regressed on an indicator variable denoting treatment status and the estimated propensity score. (included in the analysis of study)
- Not recommended for eliminating baseline differences as it does not allow balancing of covariates across treated and control groups

Propensity score stratification

- Stratifying patients into mutually exclusive subsets based on their estimated propensity score. (separates design from analysis of study)
- Patients within strata have similar (baseline) values of the propensity score.
- Not recommended for eliminating baseline differences.

Propensity score matching

- Creating matched sets of treated and untreated patients who share a similar value of the propensity score. (separates design from analysis of study)
- PSM is recommended over stratification or covariate adjustment as it eliminates greater proportion of systemic differences in baseline characteristics between treated and untreated. Not recommended for time-varying treatment or time-varying confounding.

Inverse probability weighting

- Generates a pseudo-population in which exposures are independent of confounders, enabling estimation of marginal structural model parameters. (separates design from analysis of study)
- Suitable for baseline imbalances and time-varying confounding. Not recommended when propensities are small (close to 0) as weights can be unstable.

Parametric G-formula

- Models the joint density of the observed data to generate potential outcomes under different hypothetical treatment strategies (included in the analysis of the study).
- Suitable for longitudinal data with a time-varying treatments and can adjust for time-varying confounders that are affected by prior exposures.

G-estimation

- Exploits the conditional independence between the exposure and potential outcomes to estimate structural nested model parameters (included in the analysis of the study).
- Suitable to estimate the joint effect of a sequence of treatments, when dealing with continuous exposures or when standard assumptions fail.

Sources:
•Naimi AI, Cole SR, Kennedy EH. An introduction to g methods. Int J Epidemiol. 2017 Apr 1;46(2):756-762.
•Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav Res. 2011 May;46(3):399-424.
•Hernán MA, Robins JM (2020). Causal Inference: What If (2020). Boca Raton: Chapman & Hall/CRC.

Supplement S1: Search strategy

Initial search	(time var*[tiab] OR time dependent*[tiab] AND iptw[tiab] OR inverse probability[tiab]) OR ("Propensity Score"[Mesh] OR propensity score*[tiab])	28373
Included propensity mesh to reduce the number of papers	("Propensity Score"[Mesh] OR propensity score*[tiab]) AND (time var*[tiab] OR time dependent*[tiab] OR iptw[tiab] OR inverse probability[tiab])	1566
Added risk-set matching	(risk-set matching[tiab]) OR ("Propensity Score"[Mesh] OR propensity score*[tiab]) AND (time var*[tiab] OR time dependent*[tiab] OR iptw[tiab] OR inverse probability[tiab])	1570
Added g-methods to the search and studies published before 2021	(risk-set matching[tiab]) OR ("Propensity Score"[Mesh] OR propensity score*[tiab]) OR (“g-methods”[tiab] OR “g-formula”[tiab] OR “g-estimation”[tiab] OR “parametric g-formula”[tiab]) OR (iptw[tiab] OR inverse probability[tiab]) AND (time var*[tiab] OR time dependent*[tiab] or longitudinal*[tiab]) Filters: from 1992 - 2020	2081
Added “Marginal structural Cox model” to the search	(risk-set matching[tiab]) OR ("Propensity Score"[Mesh] OR propensity score*[tiab]) OR ("g-methods"[tiab] OR "g-formula"[tiab] OR "g-estimation"[tiab] OR "parametric g-formula"[tiab]) OR (iptw[tiab] OR inverse probability[tiab] OR Marginal structural Cox model[tiab]) AND (time var*[tiab] OR time dependent*[tiab] or longitudinal*[tiab]) Filters: from 1992 - 2020	2087
Total combined		2140

Supplement S2: Empirical example details from Box 1

Empirical examples

Data from the Osteoarthritis Initiative (OAI) was used for two empirical examples. The OAI is a multicentre, longitudinal cohort study that included patients with (or at risk for) symptomatic femoral-tibial knee osteoarthritis (OA) with a follow-up up to 108 months, available for public access at https://data-archive.nimh.nih.gov/oai/. We extracted a large set of variables from the OAI that were measured at baseline and annual follow-up visits (12 to 108 months), these include general patients characteristics (age, gender, history of knee symptoms, physical activity, weight, care access), clinical variables (knee symptoms, radiographic signs of OA, hand OA), quality of life measurements (12-Item Short Form Survey (SF-12)), functional scores (Knee injury and Osteoarthritis Outcome Score (KOOS), Western Ontario and McMasters Osteoarthritis index (WOMAC)) and time-varying treatments (meniscectomy, knee replacement surgery, corticosteroid injections). Missing values were imputed through single imputation using predictive mean matching for continuous variables and logistic regression for categorical variables.

To investigate the impact of the different confounding adjustment methods on the outcome, two empirical examples with a time-varying treatment were selected that we previously published using data from the OAI: 1) the effect of meniscectomy (surgical removal of the meniscus) on the risk to receive knee replacement surgery and 2) the effect of intra-articular corticosteroid injections on the risk to receive knee replacement surgery.[19,20]

Statistical methods

In total, we compared nine methods that were the most commonly used adjustment methods found in the mapping review for both empirical examples: four methods that matched using baseline covariates, four time-dependent methods, and no matching. Confounding factors included in all eight correction methods were: patient characteristics (age, gender, BMI, physical activity, health care access, treatment centre, education, family history with OA, occupation), clinical variables (knee

medication use, hand OA at baseline, knee symptoms at baseline, radiographic confirmed OA), quality of life scores (SF-12 subscales), and functional scores (KOOS and WOMAC). After adjustment, Cox proportional hazard models were applied to estimate the treatment effect and confidence intervals.

The baseline methods consisted of PSM, IPW with a point treatment (yes/no), covariate adjustment using the propensity score, and conventional covariate adjustment (CCA) using baseline covariates and a point treatment. For PSM, the propensity score was calculated for every patient (the probability of a patient being assigned to the treatment given a set of observed covariates) and subsequently treated and control patients were matched using a 1:1 matching ratio without replacement, a caliper of 0.20 and a nearest neighbour matching algorithm, as nearest neighbour is commonly used and results in less biased estimates compared to the other matching algorithms.[21] Covariate balance was assessed by calculating the standardized mean difference (SMD) and by plotting the balance between patients and controls. Balance smaller or equal to 0.10 SMD were assumed to have appropriate balance.[2] IPW was performed to build a marginal structural model able to balance the covariates at baseline (marginal structural model with point treatment; patients were either labelled as treated or untreated). For IPW we used unbalanced weights and the weights were visually inspected. Similar to PSM, a 0.10 SMD was assumed to have an appropriate balance. Confidence intervals were estimated using 1000 bootstraps. Covariate adjustment using the propensity score was performed by calculating the propensity score using logistic regression and subsequently the propensity score was added to the Cox regression. Conventional covariate adjustment was performed by including the same set of covariates in the Cox regression without any prior adjustment.

The time-dependent methods consisted of time-dependent propensity score matching (tdPSM), IPW with time-varying treatment, parametric g-formula, and CCA with time-varying treatment and covariates.[5,15,17] Time-dependent propensity score matching was performed by sequentially

matching treated patients with all available controls at time of treatment using a 1:1 nearest neighbour matching algorithm without replacement using a caliper of 0.2. After matching a patient to a control, both were removed from the dataset to avoid further matches. Similar to the baseline methods, IPW was used to create a marginal structural model but with time-varying treatment and time-varying covariates. Likewise, we used unbalanced weights and the weights were visually inspected and balance was assessed. Confidence intervals were estimated using 1000 bootstraps.

Robins' g-formula (also known as parametric g-formula or parametric g-computation) is an alternative method to recover effects of time-varying treatment under untestable assumptions, given that sufficient covariates are measured to control for confounding by unmeasured risk factors.[22] The causal effect is measured by comparing the treatment effect between an always exposed- and a never exposed scenario. Conventional covariate adjustment with time-varying covariates and treatment was performed by including these variables in the Cox regression.

Finally, we performed one crude analysis by only including the time-varying treatment in the Cox regression. All analyses and simulations were performed using R (version 4.0.2, The R Foundation for Statistical Computing, Vienna, Austria) using packages 'mice', 'MatchIt', 'WeightIt', 'gfoRmula', 'plotly', 'coxphw', 'boot', and 'survival'.[12,22–29]

Results

In total, nine methods were compared for both empirical examples: four methods that adjust using baseline covariates (PSM, IPW using point treatment, CA using the propensity score, CCA), four time-dependent methods (tdPSM, IPW using time-varying treatment, parametric g-formula, CCA) and one without adjustments. (see figure in Box 1)

In the meniscectomy example, patients who underwent meniscectomy had an HR of 3.0 (95% CI: 1.97– 4.57), 2.42 (95% CI: 1.50 – 4.16), 2.41 (95% CI: 1.79 – 3.25), and 2.76 (95% CI: 2.03 – 3.76) to receive knee replacement surgery for PSM, IPW, CA using the propensity score, and CCA using the baseline covariates, respectively. The time-dependent strategies resulted in lower hazard ratios: HR of 2.00 (95% CI: 1.32 – 3.02), 2.05 (95% CI: 1.78 – 2.40), 2.03 (95% CI: 1.83 – 2.21) and 2.13 (95% CI: 1.62 – 2.79) for tdPSM, IPW, parametric G-formula and CCA, respectively. Without any adjustment, an HR of 3.15 (95% CI: 2.37 – 4.20) was found.

The results from intra-articular corticosteroid injection examples were more consistent between the baseline and time-dependent methods. Patients that receive intra-articular corticosteroid injections had a higher risk to receive knee replacement surgery with an HR of 1.64 (95% CI: 1.42 – 1.92), 1.53 (95% CI: 1.42 – 1.65), 1.58 (95% CI: 1.33 – 1.88), and 1.59 (95% CI: 1.36 – 1.87) for the baseline methods (PSM, IPW, CA using the propensity score, and CCA, respectively) and an HR of 1.61 (95% CI: 1.38 – 1.87), 1.49 (95% CI: 1.36 – 1.57), 1.65 (95% CI: 1.53 – 1.85) and 1.63 (95% CI: 1.39 – 1.91) for the time-dependent methods (tdPSM, IPW, parametric g-formula, CCA, respectively). No adjustment resulted in an HR of 2.12 (95% CI: 1.81 – 2.48).

Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
TITLE			
Title	1	Identify the report as a scoping review.	1
ABSTRACT			
Structured summary	2	Provide a structured summary that includes (as applicable): background, objectives, eligibility criteria, sources of evidence, charting methods, results, and conclusions that relate to the review questions and objectives.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known. Explain why the review questions/objectives lend themselves to a scoping review approach.	4
Objectives	4	Provide an explicit statement of the questions and objectives being addressed with reference to their key elements (e.g., population or participants, concepts, and context) or other relevant key elements used to conceptualize the review questions and/or objectives.	4
METHODS			
Protocol and registration	5	Indicate whether a review protocol exists; state if and where it can be accessed (e.g., a Web address); and if available, provide registration information, including the registration number.	-
Eligibility criteria	6	Specify characteristics of the sources of evidence used as eligibility criteria (e.g., years considered, language, and publication status), and provide a rationale.	5
Information sources*	7	Describe all information sources in the search (e.g., databases with dates of coverage and contact with authors to identify additional sources), as well as the date the most recent search was executed.	5
Search	8	Present the full electronic search strategy for at least 1 database, including any limits used, such that it could be repeated.	5 & Supplement S1
Selection of sources of evidence†	9	State the process for selecting sources of evidence (i.e., screening and eligibility) included in the scoping review.	5
Data charting process‡	10	Describe the methods of charting data from the included sources of evidence (e.g., calibrated forms or forms that have been tested by the team before their use, and whether data charting was done independently or in duplicate) and any processes for obtaining and confirming data from investigators.	5
Data items	11	List and define all variables for which data were sought and any assumptions and simplifications made.	5
Critical appraisal of individual sources of evidence§	12	If done, provide a rationale for conducting a critical appraisal of included sources of evidence; describe the methods used and how this information was used in any data synthesis (if appropriate).	5

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
Synthesis of results	13	Describe the methods of handling and summarizing the data that were charted.	6
RESULTS			
Selection of sources of evidence	14	Give numbers of sources of evidence screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally using a flow diagram.	8 & 9
Characteristics of sources of evidence	15	For each source of evidence, present characteristics for which data were charted and provide the citations.	8
Critical appraisal within sources of evidence	16	If done, present data on critical appraisal of included sources of evidence (see item 12).	-
Results of individual sources of evidence	17	For each included source of evidence, present the relevant data that were charted that relate to the review questions and objectives.	8
Synthesis of results	18	Summarize and/or present the charting results as they relate to the review questions and objectives.	8
DISCUSSION			
Summary of evidence	19	Summarize the main results (including an overview of concepts, themes, and types of evidence available), link to the review questions and objectives, and consider the relevance to key groups.	10
Limitations	20	Discuss the limitations of the scoping review process.	10
Conclusions	21	Provide a general interpretation of the results with respect to the review questions and objectives, as well as potential implications and/or next steps.	12
FUNDING			
Funding	22	Describe sources of funding for the included sources of evidence, as well as sources of funding for the scoping review. Describe the role of the funders of the scoping review.	13

JBI = Joanna Briggs Institute; PRISMA-ScR = Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews.
* Where *sources of evidence* (see second footnote) are compiled from, such as bibliographic databases, social media platforms, and Web sites.
† A more inclusive/heterogeneous term used to account for the different types of evidence or data sources (e.g., quantitative and/or qualitative research, expert opinion, and policy documents) that may be eligible in a scoping review as opposed to only studies. This is not to be confused with *information sources* (see first footnote).
‡ The frameworks by Arksey and O'Malley (6) and Levac and colleagues (7) and the JBI guidance (4, 5) refer to the process of data extraction in a scoping review as data charting.
§ The process of systematically examining research evidence to assess its validity, results, and relevance before using it to inform a decision. This term is used for items 12 and 19 instead of "risk of bias" (which is more applicable to systematic reviews of interventions) to include and acknowledge the various sources of evidence that may be used in a scoping review (e.g., quantitative and/or qualitative research, expert opinion, and policy document).

From: Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMAScR): Checklist and Explanation. Ann Intern Med. 2018;169:467–473. doi: 10.7326/M18-0850.